Growth, Decline, Rebirth: Quantifying Regional and Local Outcomes in the Midwest Using Principal Component Analysis (PCA), 1970-2010

Honors Thesis City & Regional Planning

Presented in Partial Fulfillment of the Requirements for *graduation with honors research distinction in City and Regional Planning* in the Knowlton School

The Ohio State University

Samuel Mattern

The Ohio State University

May 2021

Faculty Thesis Mentor: Dr. Jennifer Clark, City and Regional Planning

Table of Contents

Introduction / Overview

As cities across the United States continue to face considerable economic and social change, public officials, planners, and researchers seek to understand the dimensions of change. One of the failures of U.S. planning has been disregarding the larger multiple county regional context when analyzing cities (Gerkens, 2000). Ignoring the development of these multi-county regions, or Metropolitan Statistical Areas (MSAs), and their specific characteristics results in weak policy and planning responses. Analyzing data from an MSA perspective can provide a more comprehensive understanding of the challenges facing cities and the regions they fall within (Malecki, 2007). Addressing the gaps in knowledge about how and why cities change must remain a primary focus of planning research.

One of the regions facing considerable change is the Midwest. This region contains eleven states as defined by the Census Bureau (U.S. Census, 1995). Since the 1970s, the Midwest has experienced significant economic changes that have further impacted social outcomes. Globalization remains central to many of these changes, having a measured impact on the Midwest's industrial identity (Florida, 2016). The Midwest did not have a singular response to globalization and other economic forces. This has resulted in varying degrees of economic and social health for different localities within the region (Austin & Hitch, 2020). Despite significant research about how Midwest and its MSAs have changed since 1970, policy recommendations still fall short of providing a comprehensive solution to some of the Midwest's most significant problems (Clark & Doussard, 2019). It is still unclear why specific MSAs in the Midwest have changed and how this contributes to the identity of the Midwest as a region. This analysis seeks to highlight some future areas for research and further consideration.

This research project aims to provide a more focused analysis of the variables impacting economic and community development within the Midwest region. Deeper knowledge of the trends in regional change can inform further research and policy recommendations for Midwestern MSAs. This research project utilizes Principal Component Analysis to identify significant trends in the Midwest region's measures of economic and social health and compare individual MSA trends in comparison with one another. These objectives are pursued to provide a more comprehensive understanding of change at the regional and local level in the Midwest to better inform future research and decisions made by public officials and planners.

Literature Review

Principal Component Analysis (PCA)

This research project seeks to utilize Principal Component Analysis to identify significant trends in Midwest regional data and guide future research into development at the regional geographic scale. This method has existed for over one hundred years, first being introduced by Karl Pearson, and subsequently developed by Harold Hotelling (Hotelling, 1933; Pearson, 1901). This type of analysis is a data reduction tool that receives many inputs and produces an output of smaller components that outline which variables account for the data's variance.

Testing the validity of PCA remains a primary focus of current research involving this method. One study suggests that the number of variables included within PCA determines its usefulness and therefore "must be defined" (Wold et al., 1987, pp. 47). Outliers or unusually strong groupings of variables into components could be an indicator that the method cannot provide useful analysis, normally because of the data collected. This paper ultimately argues that PCA should not be viewed as the end goal but rather a tool utilized to guide future research.

Another study researching the quality of the PCA method determines that it holds the most validity when understood as a simplification or reduction tool (Abdi & Williams, 2010, pp. 451). It argues that PCA cannot comprehensively capture the entirety of the data but rather describe similarities between data points to reveal possible areas for further investigation. PCA not only stands alone as an analysis tool but can be utilized in conjunction with other methods. This research ultimately concludes that PCA is one of the most "versatile" methods of multivariate analysis that has existed for some time.

One of the possible limitations of PCA within the context of this research project is its validity when comparing data from multiple points in time. A recent paper outlines some of the longitudinal limitations of PCA with planning research, arguing that many of definitions of data are not "constant in space and time" (Liborio et al., 2020). This research outlines appropriate considerations to make when utilizing PCA in a longitudinal analysis. Two indicators must be constructed, with the second indicator utilizing the eigenvalues from the first PCA that is performed. A study outlining this method analyzed the Quality of Life (QOL) index for 31 provisional administrative divisions in Mainland China. This study describes how utilizing standardized values for each indicator makes temporal analysis possible (Li & Wang, 2013).

Planning Research Utilizing Principal Component Analysis

Understanding how planning research has utilized Principal Component Analysis can guide the employment of this analysis tool within the context of this research project. One paper utilized PCA to generate principal components from ten different variables that were then linearly combined to create a smart growth index for six similar sized cities in Australia and Ireland (Zhang, 2017). This analysis does not include many variables, so its results may not be significant, but it does provide a framework for utilizing PCA with urban data. Another piece of

literature utilizing PCA typified urban sprawl within Qazvin, Iran (Zebardast & Ghanooni, 2019). This analysis showcased the ability of factor analysis to abstract several factors that explain variance within the city's districts. This paper also contributed to the research design by demonstrating that variables need to be selected for analysis based on previous literature and on availability of data.

Another study within the realm of planning research utilized PCA and cluster analysis to create a typology of urban immigrant neighborhoods in a variety of Metropolitan Statistical Areas (Vicino et al., 2011). The analysis highlighted the seven components generated from the PCA and the loadings each variable had within each component. Analysis of social indicators and explanation of significant loadings helped guide the research design. A similar study concerning urban land classification in the UK's West Midlands utilized PCA to reduce the "dimensionality of the land cover dataset" (Owen et al., 2006, pp. 311). This utilization of PCA is consistent with the research design and supports the idea that it will be a useful method of data reduction.

Variable Selection

The validity and utility of Principal Component Analysis relies on the quality of data selected for analysis. This tool reduces many variables into smaller groups called components to explain variance in data. If the analysis begins with too few variables, the PCA performed will not produce any meaningful results. This limitation means that as many variables as possible must be analyzed to accomplish this research's goal of identifying significant development trends in Midwestern Metropolitan Statistical Areas. Previous development literature (especially studies analyzing the Midwest) will guide the collection of variables needed to utilize this data analysis tool. The variables used for analysis within each paper covered in the literature review

were synthesized. Appendix A lists each variable, how it is measured, and which piece of literature utilized it. The research design portion of this paper explains why certain variables were selected from the list for use within PCA.

Much of the literature on MSAs, the Midwest, and older industrial cities focuses on analyzing which dimensions of change hold meaning for researchers and public officials. A recent paper that outlines the collection of variables for analyzing MSAs guided this section of the literature review (Van Leuven & Hill, 2020). The study utilized a cluster-discriminant analysis to better classify and understand "legacy cities" (Van Leuven & Hill, 2020, pp. 1). The methodology section outlines the seventeen variables selected, why and how they were grouped together, and the reasoning that each variable holds validity within the analysis. There were five control variables, five variables that describe "legacy assets", and eight variables that describe "legacy liabilities" (Van Leuven & Hill, 2020, pp. 4-5). This paper not only provided a considerable number of variables to be considered within the analysis but also referenced numerous additional papers that were consulted.

Several of these studies specifically analyzed the United States' older industrial cities and recommended policy based on the analysis. The first paper focuses on the three concepts of growth, prosperity, and inclusion and utilizes multiple variables for each concept to describe how a specific older industrial city, county, or community performs (Berube & Murray, 2018). It utilizes these concepts to introduce the assets and challenges of older industrial cities and suggest several policy objectives. The second paper analyzing older industrial cities grouped the variables identifying these cities into two categories (Vey, 2007). The first category describes the cities' economic condition, and the second focuses on residential economic well-being. These indicators are used to identify and compare older industrial cities with the other cities in

America. A third piece of literature ranks the relative strength of eighteen legacy cities selected for analysis (Mallach & Brachman, 2013). This analysis utilized fifteen factors, both economic and social, to determine the performance of the legacy cities.

Central cities within MSAs are the focus of the next few pieces of literature. The first paper researches the income disparities between U.S. central cities and their suburbs (Hill & Wolman, 2011). This study measures the impact specific variables have on the disparity between the central cities within MSAs and the suburbs surrounding them. The second paper creates an index comprised of four variables (poverty rate, unemployment rate, change in population over the preceding decade, median household income) to identify cities with "municipal distress" (Furdell et al., 2005, pp. 283). This index was then utilized to compare these cities with central cities that were non-distressed and America as a whole. Research on economic changes within regions also focuses on entire MSAs. One paper seeks to understand the impact a range of variables have on MSA employment and gross metropolitan product (Blumenthal et al., 2009). This analysis utilizes seventeen variables to construct an analysis of MSAs within the United States.

Other pieces of literature investigate the social dimension of regions in much greater detail. One paper analyzes the validity of the Social Vulnerability Index (SoVI), a tool used to calculate social vulnerability to natural hazards at the county level in the United States (Schmidtlein et al., 2008). This study examines how sensitive specific quantitative elements of the index responded to changes in its construction, geography of analysis, and the variables used within the analysis. The original index was comprised of twenty-six variables measured at the county level.

Research Design

Description and Study Area

This research study is a descriptive and exploratory analysis of the change in economic and social variables within Midwestern Metropolitan Statistical Areas (MSAs). The specific geography studied is the East North Central Division of the Midwest region. As defined by the U.S. Census Bureau, this division includes Indiana, Illinois, Michigan, Ohio, and Wisconsin (U.S. Census, 1984). The selection of this specific division within the Midwest region was made with several considerations. First, the amount of data that will be collected and analyzed is significant, so limiting the geographic scope will make this process achievable in an academic semester. Second, this area of the Midwest includes my hometown, thus I have a personal preference for wanting to analyze this area. Third, Principal Component Analysis may not hold as useful with a larger geographic area because of greater total variance. Analyzing a division of the Midwest selects an area that is generally more homogenous.

Variable Selection and Data Collection

Variable selection holds the most importance within the research design. As was stated in the literature review, Principal Component Analysis relies on a larger number of variables to hold more statistical significance. Selecting variables that appropriately represent some of the changes this project seeks to describe relied on (1) collecting variables from previous literature and (2) subsequently reducing the number of variables to be analyzed based on theoretical concerns. The collection of variables is outlined in Appendix A. The first removal of variables occurred for those already including a component of time. Several pieces of literature presented variables as a part of their arguments that seek to measure change over time. As measuring

change over time is also an element of this research project, these variables were not selected for analysis.

Another removal occurred for categorical variables. According to Starkweather (2010), PCA is not normally used in a "confirmatory" manner, but rather an exploratory one (pp. 3). He describes that traditional PCA should be utilized when trying to reduce data while accounting for maximum total variance. Because categorical data is dichotomous, this type of variable was not selected for analysis.

The last removal of variables was dependent on the availability of data. Data was collected from the years 1970 and 2010. Due to the changing nature of data collection and research at the Metropolitan Statistical Area geography level, there were variables that cannot be analyzed at both time periods. 27 distinct variables' data was collected for the 228 counties comprising the 67 MSAs in the study area for both 1970 and 2010. The final list of variables is outlined in Appendix B.

Description of Two-Tiered Analysis

This analysis is separated into two tiers. The first tier of analysis will investigate change in the amount of variance that specific variables describe in the Midwest East North Central region. The second tier of analysis investigates the relative change of a specific MSA's component score in relation to the other MSAs in the region. Socioeconomic data from 1970 and 2010 is utilized for this two-tiered analysis. Data agglomeration takes place within Microsoft Excel, while IBM's SPSS software is utilized for statistical calculations.

Analysis / Findings

First-Tier Analysis – Change in Variance Described at Regional Level

As described in the research design, the first tier of analysis focuses on the amount of variance described by specific variables in the Midwest East North Central region. This approach for temporally comparing variance described at the regional level follows the method outlined in the literature review. Four tests for validity are performed on each data set. If these assumptions are met, two Principal Component Analyses are performed, one on data from 1970 and one on data from 2010. The components generated as well as variable loadings on components are then analyzed.

Before testing the data, all variables were standardized. Four tests were then performed on the data: checking for a linear relationship between variables, checking for sampling adequacy (KMO Measure of Sampling Adequacy), checking that the data is suitable for reduction (Bartlett's test of sphericity), and checking that there are no significant outliers. For the data set from 1970, there was a linear relationship between variables, the KMO measure was 0.673, Bartlett's test yielded a value of < 0.001, and significant outliers were then removed from the data set. For the data set from 2010, there was a linear relationship between variables, and significant outliers were removed, but the KMO measure and Bartlett's test measure did not have an output. The PCA still outputted from this dataset, but this means that there are linear dependencies between variables.

As the same variables were utilized for both data sets, this outcome could mean that certain variables became more closely tied over the four-decade period in which they were measured. Comparison of the variable loadings for each PCA reveals which variables are now more linearly related. Appendix C outlines some of the methodological decisions for Tier 1.

First PCA – 1970 Data

The first PCA was performed on the data set from 1970 which included 27 variables for 67 distinct Metropolitan Statistical Areas (MSAs) in the study area. Eight components were produced explaining approximately 84% of the variance in the data, with the first three components each accounting for at least 10% of the variance in the data (17.3%, 13.1%, and 12.4% respectively). Table 1 below shows each of the components, eigenvalues, and percentage of variance.

Table 1

Variance Explained by First PCA[a] – 1970 Data

Component	Rotation Sums of Squared Loadings		
	Eigenvalue	Percentage of Variance	Cumulative Percentage
1	4.673	17.308	17.308
2	3.530	13.074	30.382
3	3.352	12.416	42.798
4	2.540	9.407	52.205
5	2.463	9.123	61.328
6	2.252	8.342	69.670
7	2.157	7.989	77.658
8	1.938	7.177	84.835

[a]Rotated using Varimax with Kaiser normalization in 22 iterations.

Understanding the meaning of each component requires investigating how closely each variable correlates with the eight produced components. The loadings of each variable describe the level of correlation with the specific component, with values closer to 1.0 or -1.0 indicating a stronger correlation. Any loading with an absolute value of 0.45 or higher was kept in this analysis. Table 2 below shows the Rotated Component Matrix with variable loadings for each

component. This table also includes the communality for each variable, which is the variance in a variable that is explained by the eight produced components.

Table 2

Rotated Component Matrix with Communalities – 1970 Data

Component	1	2	3	4	5	6	7	8	Commun-alities
MSA-Wide Population Density per square mile					0.50			0.48	0.91
Percent over age 65	-0.56								0.80
Percent under the age of 18	0.77								0.81
Civilian labor force participation rate	0.83								0.86
Vacancy of housing units				0.90					0.91
Median home value ($) of owner-occupied units (2010 dollars)					0.54				0.85
Poverty rate				0.80					0.91
Percentage of population age 25 and older with bachelor's degree		0.87							0.89
Percentage of population age 25 and older with less than high school		-0.79							0.84
Jobs					0.55		-0.59		0.86
Per-capita income (2010 dollars)		0.48							0.82
Median household income (2010 dollars)		0.61							0.97
Unemployment rate						-0.74			0.84
Percent foreign-born population					0.87				0.88
Percentage of Black residents			0.87						0.86
Percentage of female participation in civilian labor force	0.88								0.94
Percentage of female-headed households			0.89						0.91
Median gross rent ($) of renter-occupied units (2010 dollars)		0.77							0.93
Percentage of population under 5	0.88								0.89
Percentage of institutionalized elderly population	-0.64								0.66
Average number of people per household						-0.67			0.87
Percentage of renter-occupied housing units						0.74			0.80
Percentage of employment in primary industry: farming, fishing, mining, or forestry							0.63		0.81
Percentage employed in transportation, communications, or other public utilities							-0.84		0.88
Percentage of population living in urban areas				-0.46				0.52	0.72
Percentage of females								0.80	0.78
Percentage of Hispanic persons			0.80						0.69

Variables with high loadings on Component 1 relate to the age and labor force composition of each MSA. This component indicates MSAs with a significant youth population, a relatively small elderly population (especially those residing in institutions), high civilian labor-force participation, and high female participation in the civilian labor force. This could indicate MSAs with a large working population, many of these workers with children, with the female partner in a heterosexual relationship participating in the work force nearly as often the male partner.

Above-average income and education typify the variables with high loadings for Component 2. This component describes MSAs with a highly educated population (at least bachelor's degree), a smaller low-educated population (less than high school education), modest per-capita income, high median household income, and high rent for renter-occupied units. This could indicate MSAs with many college graduates and very few people with less than high school diploma. These people earn a relatively high income that can pay for their above-average rent.

High minority populations and female-headed households describe the variables with high loadings for Component 3. This component describes MSAs with large Black and Hispanic populations, as well as many female-headed households. This could indicate MSAs with high minority population (as many were not tracked on this census). The difficulty of immigrating could lead to many non-traditional households with female heads.

Variables with high loadings on Component 4 describe blight and decay within the MSA. This component describes MSAs with high vacancy rates, high poverty rates, and interestingly, a smaller urban population. This could be indicative of higher suburban populations that have vacated homes and taken economic and educational opportunity with them.

The variables with high loadings for Component 5 do not seem to be grouped in a specific type of variable. This component describes MSAs with decent density per square mile, modest home value of owner-occupied housing units, more jobs, and a higher foreign-born population. This could indicate MSAs that have more diverse employment opportunities that support immigrants. These immigrants own their own homes in a relatively dense multi-county regional framework.

Unemployment rate, average number of people per household, and percentage of renter-occupied housing units are the variables with high loadings for Component 6. This could indicate MSAs with many single-person households that are employed and live close to where they work.

Variables with high loadings on Component 7 are the two variables that measure sectors of employment. The loading is positive and high for the agricultural sector and negative and even higher for the transportation sector. This could indicate MSAs that have a larger agricultural sector.

Component 8 has variables with high loadings that typify a dense urban MSA with an above-average female population. This could indicate MSAs that are more progressive and offer more opportunity for single women.

The first PCA demonstrates how age, income, education, the built environment, race, density and land-use, and sex impact how different MSAs are typified. The most strongly correlated variables were in the third and fourth components which described high minority populations, female-headed households, and economic and social decay in the urban centers of MSAs. The next part of the analysis focuses on generating a PCA for the data from 2010 and analyzing its nuances.

Second PCA – 2010 Data

The second PCA was performed on the data set from 2010 which included 27 variables for 67 distinct Metropolitan Statistical Areas (MSAs) in the study area. Six components were produced explaining approximately 83% of the variance in the data, with the first component accounting for almost 35% of the variance in data, and each of the next three accounting for at least 10% (13.7%, 12.4%, and 10.3% respectively). Table 3 below shows each of the components, eigenvalues, and percentage of variance.

Table 3

Variance Explained by Second PCA – 2010 Data

Component	Rotation Sums of Squared Loadings		
	Eigenvalue	Percentage of variance	Cumulative percentage
1	9.432	34.933	34.933
2	3.712	13.748	48.681
3	3.342	12.378	61.059
4	2.780	10.296	71.356
5	1.781	6.595	77.951
6	1.448	5.363	83.314

[a]Rotated using Varimax with Kaiser normalization in 7 iterations.

To understand the meaning of each component, variable correlations with each of the six generated components must be investigated. The loadings of each variable describe the level of correlation with the specific component, with values closer to 1.0 or -1.0 indicating a stronger correlation. Any loading with an absolute value of 0.5 or higher was kept. Table 4 below shows the Rotated Component Matrix with variable loadings for each component. This table also includes the communality for each variable, which is the variance in a variable that is explained by the six produced components.

Table 4

Rotated Component Matrix with Communalities – 2010 Data

Component	1	2	3	4	5	6	Communalities
MSA-Wide Population Density per square mile		0.88					0.81
Percent over age 65	-0.63			-0.57			0.86
Percent under the age of 18			0.95				0.95
Civilian labor force participation rate	0.88						0.94
Vacancy of housing units	-0.59					-0.57	0.75
Median home value ($) of owner-occupied units (2010 dollars)	0.84						0.93
Poverty rate	-0.91						0.89
Percentage of population age 25 and older with bachelor's degree	0.68						0.80
Percentage of population age 25 and older with less than high school	-0.81						0.83
Jobs		0.51					0.44
Per-capita income (2010 dollars)	0.78						0.83
Median household income (2010 dollars)	0.93						0.93
Unemployment rate	-0.79						0.90
Percent foreign-born population				0.76			0.78
Percentage of Black residents	-0.59	0.61					0.77
Percentage of female participation in civilian labor force	0.89						0.91
Percentage of female-headed households	-0.79	0.50					0.91
Median gross rent ($) of renter-occupied units (2010 dollars)					-0.50		0.76
Percentage of population under 5 years			0.81				0.91
Percentage of institutionalized elderly population					0.86		0.77
Average number of people per household			0.86				0.88
Percentage of renter-occupied housing units			-0.62	0.64			0.94
Percentage of employment in primary industry: farming, fishing, mining, or forestry		-0.75					0.70
Percentage employed in transportation, communications, or other public utilities	0.89						0.91
Percentage of population living in urban areas		0.77					0.84
Percentage of females		0.56				0.60	0.76
Percentage of Hispanic persons				0.72			0.78

Component 1 has many variables with high loadings that describe the education, income, employment, and housing of the MSAs. This component describes MSAs with high civilian labor force participation rates (especially females) and low unemployment rates, low vacancy rates and high median home value, low poverty rates, high per-capita income and median household income, smaller Black population, and more people employed in transportation, communication, and other public utilities. This component has so many variables with high loadings, suggesting there may be many MSAs that have similar measures for all these variables.

Variables with high loadings for Component 2 focus on females, dense urban areas, little agriculture, more Black residents, and a decent number of jobs. This could suggest that these MSAs have a dense urban core with a higher Black population, more female-headed households, and service-oriented jobs. Larger-than-average families are the focus of variables with high loadings for Component 3. This could typify less dense, suburban MSAs with a higher number of single-family homes with larger than average family size. Component 4 has variables with high loadings that describe a high immigrant population, a younger adult population (not many over age 65), and high renter population. This could indicate MSAs with more first-generation immigrant neighborhoods, especially Hispanic ones.

Comparison of First and Second PCAs

The first and second PCAs indicate remarkable change in the Midwest region over the forty-year difference in data. Based on the difference in number of components generated (8 and 6) and the number of variables presenting high loadings for each component (average of 3.5 and 5.5 per component), it seems that the data for 2010 is more homogenous than the data from 1970. This could indicate that MSAs within the Midwest region experienced similar challenges

between 1970 and 2010 leading to less varied outcomes for variables measuring socioeconomic performance.

It is also possible that trends extending further back than this forty-year period continued to manifest. One of these trends is segregation within the built environment. Components 3 and 5 from the first PCA describe high Black and immigrant populations, respectively. These are like Components 2 and 4 from the second PCA, which typify dense urban areas with Black residents and high immigrant renter populations. Although the loadings were not much stronger for the Black population component, the immigrant population's component became much more correlated, suggesting even greater spatial segregation into immigrant neighborhoods.

Suburbanization could possibly be another trend beginning before 1970 that continued to develop between 1970 and 2010. Component 4 in the first PCA describes a vacated urban core of MSAs with high poverty rates. This could relate in part to the generation of Component 3 from the second PCA, which describes larger-than-average families that do not rent. If trends like spatial segregation and suburbanization are viewed through a larger temporal lens, it is possible to see how the components generated for 1970 and 2010 fit within a larger evolution at the multi-county regional scale.

Second-Tier Analysis – Change in Component Scores for MSAs

As described in the research design, the second tier of analysis investigates the change of an MSA's component score in relation to the other MSAs in the Midwest East North Central region. This process expands on the first tier of analysis by providing more clarity on how individual MSAs have changed over time and in relation to each other. The first part of analysis is identical. After the data is inputted into the software, four tests will be conducted on the data: checking for a linear relationship between variables, checking for sampling adequacy, checking

that the data is suitable for reduction, and checking that there are no significant outliers. If these assumptions are met, two Principal Component Analyses are performed, and principal components are generated for 1970 and 2010 data.

Before testing the data, all variables were standardized. Four tests were then performed on the data: checking for a linear relationship between variables, checking for sampling adequacy (KMO Measure of Sampling Adequacy), checking that the data is suitable for reduction (Bartlett's test of sphericity), and checking that there are no significant outliers. For the data set from 1970, there was a linear relationship between variables, the KMO measure was 0.673, Bartlett's test yielded a value of < 0.001, and significant outliers were then removed from the data set. For the data set from 2010, there was a linear relationship between variables, and significant outliers were removed, but the KMO measure and Bartlett's test measure did not have an output. The PCA still outputted from this dataset, but this means that there are linear dependencies between variables.

Analysis of individual MSAs relies on the generation of component scores for each principal component generated. According to the study outlined in the literature review, temporal analysis is possible if the eigenvalues or "loadings" from one PCA are utilized to generate component scores for both sets of data. Component scores were generated normally for the data from 1970 by multiplying variable loadings by the standardized data z-scores for every variable within each principal component and summing the totals. This process was replicated for the data from 2010 but utilized the loadings from the First PCA (1970 Data) in combination with standardized 2010 data z-scores.

The strongest five correlations for each principal component are analyzed first, followed by an analysis of the five largest changes in component scores for each component between 1970 and 2010. Appendix C outlines some of the methodological decisions for Tier 2.

Table 5

Largest Component Scores – 1970 Data

Component	1	2	3	4	5	6	7	8
1	Columbus, IN (5.71)	Kalamazoo-Portage, MI (6.29)	Detroit-Warren-Dearborn, MI (3.90)	Niles, MI (3.87)	Youngstown-Warren-Boardman, OH-PA (4.16)	Muskegon, MI (-3.60)	Toledo, OH (-2.87)	Duluth, MN-WI (-3.32)
2	Danville, IL (-5.31)	Wheeling, WV-OH (-6.21)	Chicago-Naperville-Elgin, IL-IN-WI (3.88)	Bloomington, IN (3.49)	Racine, WI (3.59)	Madison, WI (3.05)	Kokomo, IN (2.81)	Carbondale-Marion, IL (-2.85)
3	Lansing-East Lansing, MI (5.29)	Huntington-Ashland, WV-KY-OH (-5.22)	Wausau-Weston, WI (-3.88)	Terre Haute, IN (3.48)	Akron, OH (3.48)	La Crosse-Onalaska, WI-MN (3.00)	Columbus, IN (2.34)	Lafayette-West Lafayette, IN (-2.56)
4	Rockford, IL (5.29)	Lansing-East Lansing, MI (4.96)	Muskegon, MI (3.57)	Duluth, MN-WI (3.03)	Toledo, OH (3.25)	Wausau-Weston, WI (-2.95)	Sheboygan, WI (2.14)	Wausau-Weston, WI (-2.43)
5	Flint, MI (5.04)	Weirton-Steubenville, WV-OH (-4.85)	Columbus, IN (-3.48)	La Crosse-Onalaska, WI-MN (-3.03)	Rockford, IL (2.90)	Duluth, MN-WI (-2.89)	Elkhart-Goshen, IN (2.12)	La Crosse-Onalaska, WI-MN (2.29)

The data from 1970 revealed that several MSAs have one of the five strongest component scores for multiple components. This could mean that these MSAs have data patterns that are more abnormal than other MSAs in the region. Columbus, IN has the strongest positive Component 1 score, a negative Component 3 score, and a positive Component 7 score. This could suggest that Columbus, IN has a large agricultural working population with children that has a smaller minority population and less female-headed households. La Crosse-Onalaska, WI-MN has a negative Component 4 score and a positive score for Components 6 and 8. This could

indicate La Crosse-Onalaska, WI-MN has a dense urban population with many renter, low vacancy rates, and more people per household.

Duluth, MN-WI's data could indicate, interestingly, the exact opposite from La Crosse-Onalaska, WI-MN. Duluth, MN-WI has a positive Component 4 score and a negative score for Components 6 and 8. This could indicate that Duluth, MN-WI has a small urban population, high vacancy and poverty rates, and less people per household. These groupings of components where scores follow each other could suggest that the variables in one component affect the measures of those in another (i.e., Components 6 and 8). Wausau-Weston, WI also has a negative correlation for both Components 6 and 8. Although it is not one of the strongest correlations, Wausau-Weston, WI also has positive correlation for Component 4, like Duluth, MN-WI.

Table 6

Largest Component Scores – 2010 Data

Component	1	2	3	4	5	6	7	8
1	Wheeling, WV-OH (-8.02)	Madison, WI (6.20)	Saginaw, MI (3.91)	Appleton, WI (-3.41)	La Crosse-Onalaska, WI-MN (3.03)	Madison, WI (3.01)	Danville, IL (2.30)	Duluth, MN-WI (-2.81)
2	Weirton-Steubenville, WV-OH (-7.87)	La Crosse-Onalaska, WI-MN (5.75)	Flint, MI (3.54)	Carbondale-Marion, IL (3.01)	Wheeling, WV-OH (-2.34)	Flint, MI (-2.39)	Terre Haute, IN (2.22)	Jackson, MI (-2.78)
3	Youngstown-Warren-Boardman, OH-PA (-6.07)	Bloomington, IL (5.49)	Wausau-Weston, WI (-3.46)	La Crosse-Onalaska, WI-MN (-2.85)	Weirton-Steubenville, WV-OH (-2.33)	Detroit-Warren-Dearborn, MI (-2.27)	Weirton-Steubenville, WV-OH (2.12)	Carbondale-Marion, IL (-2.69)
4	Minneapolis-St. Paul-Bloomington, MN-WI (5.65)	Minneapolis-St. Paul-Bloomington, MN-WI (5.03)	Detroit-Warren-Dearborn, MI (3.29)	Bloomington, IL (-2.81)	Danville, IL (-2.28)	Oshkosh-Neenah, WI (2.20)	La Crosse-Onalaska, WI-MN (-1.97)	Terre Haute, IN (-2.46)
5	Appleton, WI (5.49)	Huntington-Ashland, WV-KY-OH (-4.70)	Eau Claire, WI (-3.28)	Madison, WI (-2.78)	Terre Haute, IN (-2.18)	Grand Rapids-Kentwood, MI (-2.18)	Decatur, IL (1.94)	Wausau-Weston, WI (-2.43)

The same types of relationships between components do not seem to be present in the data from 2010. Wheeling, WV-OH has a very negative Component 1 score and a negative Component 5 score. This could suggest Wheeling, WV-OH has a smaller labor-force, with a high elderly population, lower density, and less immigrants. La Crosse-Onalaska, WI-MN again has several strong scores but for different components than the 1970 data. This time the MSA has negative scores for Components 4 and 7 and a positive score for Components 2 and 5.

Similar patterns do not exist to the same degree in the data from 2010 that were present in the data from 1970. Bloomington, IL, like La Crosse-Onalaska, WI-MN, also has a positive score for Component 2 and a negative score for Component 4 but does not have positive score for Component 5. This lack of patterns in the data could indicate more homogeneity amongst the MSAs. It could also indicate that analyzing the data from 2010 utilizing components generated with the data from 1970 may conflate the analysis between MSAs.

Table 7

Largest Change in Component Scores from 1970 to 2010

Component	1	2	3	4	5	6	7	8
1	Duluth, MN-WI (+7.70)	Madison, WI (+4.81)	South Bend-Mishawaka, IN-MI (+2.89)	Bloomington, IL (-4.93)	Youngstown-Warren-Boardman, OH-PA (-5.65)	Wausau-Weston, WI (+2.82)	Decatur, IL (+3.91)	Terre Haute, IN (-2.76)
2	Eau Claire, WI (+7.61)	La Crosse-Onalaska, WI-MN (+4.43)	Janesville-Beloit, WI (+2.68)	Bloomington, IN (-3.40)	Weirton-Steubenville, WV-OH (-3.64)	Rockford, IL (-2.58)	Terre Haute, IN (+3.90)	Wheeling, WV-OH (-2.48)
3	Kokomo, IN (-6.94)	Kalamazoo-Portage, MI (-4.00)	Kankakee, IL (+2.59)	Flint, MI (+3.33)	Akron, OH (-3.63)	Kankakee, IL (-2.44)	Carbondale-Marion, IL (+3.02)	Lima, OH (-1.84)
4	Midland, MI (-6.25)	Champaign-Urbana, IL (+3.79)	Columbus, IN (+2.14)	Jackson, MI (+3.31)	Toledo, OH (-3.56)	Eau Claire, WI (+2.32)	Wheeling, WV-OH (+2.96)	Flint, MI (+1.84)
5	Weirton-Steubenville, WV-OH (-6.14)	Wausau-Weston, WI (+3.66)	Muncie, IN (-2.00)	Eau Claire, WI (-3.17)	Columbus, IN (+2.76)	Lansing-East Lansing, MI (+2.20)	Elkhart-Goshen, IN (-2.55)	Detroit-Warren-Dearborn, MI (+1.71)

Table 7 shows the largest changes in component scores from 1970 to 2010 for each component. It is important to note that a positive or negative change does not indicate a trend further into a positive or negative component score. For example, although Columbus, IN had a positive change of 2.14 in its score for Component 7 from 1970 to 2010, it still has a negative Component 7 score in 2010 (-1.33). For this analysis, the implications of the largest change for each component are discussed.

Duluth, MN-WI experienced the largest change in its Component 1 score. This could indicate a demographic shift towards a more youthful population, more workforce opportunity, and more opportunities for women in the workforce. This trend likely means that MSAs experiencing a large shift in their score for Component 1 have more families where both partners participate in the workforce in some capacity. This leads to higher income which can support more children. Madison, WI experienced the largest change in its Component 2 score. This could indicate an increase in secondary and post-secondary education, increase in per-capita and household income, and higher average rent. As the University of Wisconsin's main campus is in Madison, this increase in education, income, and rent could be explained by higher education's expansion in the second half of the 20th century (Schofer & Meyer, 2005, pp. 908).

The largest change in Component 3 score occurred in South Bend-Mishawaka, IN-MI. This change suggests a larger minority population with an increase in female-headed households. MSAs experiencing a shift like South Bend-Mishawaka's probably saw an increase in the Black and Hispanic populations as well as other minority populations. The prevalence of female-headed households could possibly be explained the economic difficulties communities comprised of minorities typically face. Bloomington, IL saw the largest change in its Component 4 score, with its change being the first of the largest that was negative. This decrease likely indicates a

decrease in vacancy and poverty rates and an increase in the MSA's urban population. If an MSA experienced a shift like Bloomington's, this could indicate it may have received reinvestment in its urban core, leading to less blight and possible gentrification.

Youngstown-Warren-Boardman, OH-PA saw its Component 5 score change more than other MSA. Youngstown-Warren-Boardman likely experienced a decrease in density, falling property values, less jobs, and less immigration from 1970 to 2010. MSAs with a similar decrease in their Component 5 score (Weirton-Steubenville, WV-OH, Akron, OH, and Toledo, OH) may have an economic history like Youngstown-Warren-Boardman, with this decrease indicating the effect of manufacturing's decline in these areas. The largest change in the score for Component 6 occurred in Wausau-Weston, WI. This could mean Wausau-Weston experienced lower unemployment, less people per household, and a higher percentage of renter-occupied units from 1970 to 2010. MSAs with a similar increase likely experienced an increase in opportunities for new professionals with small households.

Decatur, IL experienced the largest Component 7 score change from 1970 to 2010. This increase could indicate a decrease in the number of jobs, with more employment coming from a primary industry rather transportation, communication, or another public utility. The MSA experiencing the largest change in its score for Component 8 was Terra Haute, IN. This could indicate the amount of workforce opportunities for females decreased, resulting in a greater loss of the female population and an overall decrease in density resulting from the smaller urban core.

Limitations

Although the literature review revealed a method to utilize Principal Component Analysis with several years- of data, the dimension of time adds a complication to the analysis. There is a possibility that this method does not produce meaningful results. The second tier of analysis may

be abandoned or exchanged for another analysis method in a future research study in pursuit of a more comprehensive investigation of the first tier's PCA generations.

Another limitation is the changing geography of Metropolitan Statistical Areas. According to the U.S. Census Bureau, an MSA is "a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core" (U.S. Census, 2020). This definition would suggest that the MSA geography definition accounts for regional patterns of change in economic and social organization, but this could conflate different explanations for variation in change with the PCA's selected variables.

PCA's exclusion of categorical variables may also be a limitation of this research study. By eliminating categorical variables, it is possible that variance at the regional level and between individual MSAs is not accounted for in the analysis. If this study were to be expanded upon at a later point in time, finding a different tool of analysis that can incorporate categorical variables could be an objective.

One of the stipulations of the PCA process is to remove outliers from the data set before performing the analysis. In the context of this project, limiting geography meant that specific variables (i.e., MSA-wide population density) that had large values in the biggest MSAs were considered outliers and removed. This removal of these "outliers" skews the standardization of the data and may weaken the usefulness of the analysis.

Another challenge that occurred during analysis was the second PCA's correlation matrix not being positive definite. This means that there were linear dependencies between variables. There was still an output for the PCA, which means that analysis is still useful, but may not be as valid as the first PCA utilizing data from 1970. Modifying the types of data collected for a future

study could remove the linear dependencies between variables and lead to a more comprehensive and useful analysis of regional change.

Discussion

This research project has been an incredible first experience in planning research. It has been rewarding to learn how to design a research process, collect data, perform analysis, and revise findings. Although this project provides interesting opportunities for revision and expansion of the PCA method in planning research, the process itself provided a fantastic opportunity to understand the power and scope of research. Expanding upon this research would likely rely on several changes to the research design and length. Principal Component Analysis proved a useful tool for planning research that should be utilized in more studies. This project in combination with other methods could eventually be utilized to produce an index measuring an "Metropolitan Statistical Area Performance Index" based on the variance explained by components and the weight of specific variables on individual components.

Research Design

This project investigated the East North Central division of the Midwest region as defined by the U.S. Census Bureau. As discussed in the analysis, one limitation was the removal of outliers from each data set. Many of the outliers were values from the MSAs with the largest populations (i.e., Chicago-Naperville-Elgin, IL-IN-WI, Cleveland-Elyria, OH, and Detroit-Warren-Dearborn, MI). By expanding the geographic scale of analysis to the entire Midwest region or the entire United States, the values for specific variables in these MSAs may no longer be considered outliers by analysis software. The inclusion of these values could provide more meaningful results utilizing the same analysis method.

The temporal component in this analysis should be compared to other methods of temporal analysis in a future research study. Component scores can be calculated using several different procedures. The study from Libório et al. outlines three separate methods for producing these scores. This project utilized the loadings from one PCA on another PCA's data (structured indicator). Another method involves utilizing the loadings from each PCA to produce component scores (double indicator). The third method involves combining the data from multiple years and performing one PCA. The data is then separated, and the loadings are utilized to construct the component scores (single indicator). A comparative analysis of which procedure produces the most meaningful results for this type of analysis should be an objective if this study is reproduced.

Several of the studies outlined in the literature review utilize cluster analysis to enhance the primary analysis method. Van Leuven and Hill (2020) employed cluster-discriminant analysis grouped similar MSAs and then analyzed why they were grouped together. Vicino et al. (2011) outline a method where PCA is first utilized to identify principal components in urban immigrant neighborhoods. These neighborhoods are then grouped into clusters using k-means clustering. A future iteration of this research project should utilize a cluster analysis to provide a more robust analysis of variation at the regional level beyond component scores, or correlation with components. This analysis would help expand upon the trends revealed the analysis section, including a correlation between several components.

Principal Component Analysis in Planning Research and Policy

During the literature review process, it became clear multi-county regional analysis, especially those regions in the Midwest, focused much more on economic analysis, rather than utilizing social and economic indicators to understand overall variance and change. Investigating

different analysis methods revealed the usefulness of Principal Component Analysis and its possible applications for planning research. Although the literature review outlines several other studies that employ PCA in the realm of planning, this tool could become an asset for researchers and policy experts as the universe of data measuring social and economic indicators continue to grow. As a limiting factor of the analysis was simply the smaller number of data variables from 1970, PCA's usefulness will only increase as research wades through more data over the next few decades.

Construction of MSA Performance Index

In its infancy, this research project was focused on creating a performance index for MSAs. This index would allow local planners and officials to input data for their region and receive a score, like the Human Development Index and other measures of social and economic well-being. The process of creating an index had too many considerations and methodological steps for an undergraduate thesis project, but it is a tool that this research could help create.

This tool would prompt users to input data for specific variables for their cities. It would then output a cumulative MSA Performance Score in addition to scores for each of the elements comprising the score. This would be accompanied by a short analysis that aims to explains how individual scores impact the cumulative score and how the city should interpret its weaker scores. This tool will not suggest specific policy objectives, but rather aim to demonstrate which areas the city is failing to provide public services.

Conclusion

The main objectives of this research project were to investigate the variables measuring economic and community development in the Midwest and to utilize Principal Component Analysis to describe local and regional trends in data. Greater understanding of change at local

and regional level could be utilized to inform planning and policy decisions and to predict how trends impact the future of MSAs in the region. Analyzing this study's success in meeting these objectives provides perspective on future research goals and the methods utilized.

The most useful element of the analysis was the grouping of variables into components. This process revealed which variables impact on another in a specific data set. Several groups of variables, or components, were consistent between the two data sets. This is significant because these components can be measured and observed utilizing more points in time to predict an MSA's change more accurately over time. This is also impactful because it reveals which groups of variables explain more variance in an MSA. Planning professionals often prioritize specific measures of social and economic health as being the "best" indicators of an MSA's performance. This study could give planners an empirical approach to measuring performance and creating specific policy objectives.

The second objective of this study was to observe local and regional change over time in the East North Central portion of the Midwest region. The first tier of analysis described regional change, while the second tier focused on local change. The first PCA generated eight components and the second PCA generated six components. Several of the components between the PCAs were comprised of similar variables, indicating regional consistencies in data forty years apart. The five largest changes in component scores were also measured, providing some context for which MSAs experienced the largest change from 1970 to 2010.

Perhaps the most important finding from this study was the data's convergence from 1970 to 2010. This convergence appears in the formation of a matrix that is not positive definite, meaning there are linear dependencies between variables, as well as a smaller number of principal components being generated. This finding could indicate that individual MSAs in this

region are becoming more homogenous. This finding challenges the prevailing literature on economic development, which heavily supports specialization (Fagerberg & Srholec, 2017; Kemeny & Storper, 2015). If MSAs in this region are becoming more homogenous, focusing on prioritizing specialized industries may in fact destabilize the regional economy and make its businesses less competitive.

Both objectives of this research study were successfully achieved. The Discussion section outlined several improvements that could be made for future iterations of this study, including changing the geography, restructuring the temporal component, and providing cluster analysis. Although these future iterations should prove even more useful, this study proves that grouping variables into components is not only significant but also useful in describing MSAs and observing their change over time.

References

Abdi, H., & Williams, L. (2010). Principal component analysis. *Computational Statistics, 4*(2), 433-459. https://doi.org/10.1002/wics.101

Austin, J., & Hitch, A. (2020). A vital Midwest: the path to a new prosperity. *The Chicago Council on Global Affairs.*

Berube, A., & Murray, C. (2018). Renewing America's economic promise through older industrial cities. *The Brookings Institution*.

Blumenthal, P., Wolman, H., & Hill, E. (2009). Understanding the Economic Performance of Metropolitan Areas in the United States. *Urban Studies, 46*(3), 605-627. https://doi.org/10.1177%2F0042098008100997

Bureau of Economic Analysis (2009). *Regional Economic Information System* (1967-2007) [Data set]. Bureau of Economic Analysis, U.S. Department of Commerce. CD-ROM.

Bureau of Labor Statistics. (2021). *CPI for All Urban Consumers (CPI-U)* [Data set]. Bureau of Labor Statistics. https://data.bls.gov/cgi-bin/surveymost

Clark, J., & Doussard, M. (2019). Devolution, disinvestment and uneven development: US industrial policy and evolution of the national network for manufacturing innovation. *Cambridge Journal of Regions, Economy and Society 2019*(12), 251–270. https://doi.org/10.1093/cjres/rsz009

Fagerberg, J., & Srholec, M. (2017). Explaining regional economic performance: the role of competitiveness, specialization and capabilities. *Handbook of Regions and Competitiveness,* 117-135. https://doi.org/10.4337/9781783475018.00010

Florida, R. (2016). Regional creative destruction: production organization, globalization, and the economic transformation of the midwest. *Economic Geography, 72*(3), 314-334. https://doi.org/10.2307/144403

Furdell, K., Wolman, H., & Hill, E. (2005). Did Central Cities Come Back? Which Ones, How Far, and Why?. *Journal of Urban Affairs, 27*(3), 283-305. https://doi.org/10.1111/j.0735-2166.2005.00237.x

Gerkens, L. (2000). Ten failures that shaped the 20th century American city. *Planning Commissioners Journal, Spring 2000* (38), 3a-9a.

Hill, E., & Wolman, H. (2011). Accounting for the Change in Income Disparities Between US Central Cities and Their Suburbs from 1980 to 1990. *Urban Studies, 34*(1), 43-60.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *24*(6 & 7), 417–441 & 498–520.

Kemeny, T., & Storper, M. (2015). Is Specialization Good for Regional Economic Development?. *Regional Studies, 49*(6), 1003-1018. https://doi.org/10.1080/00343404.2014.899691

Li, Z., Wang, P. (2013). Comprehensive Evaluation of the Objective Quality of Life of Chinese Residents: 2006 to 2009. *Soc Indic Res 113,* 1075–1090. https://doi.org/10.1007/s11205-012-0128-3

Libório, M.P., da Silva Martinuci, O., Machado, A.M.C., Machado-Coelho, T.M., Laudares, S., & Bernandes, P. (2020). Principal component analysis applied to multidimensional social indicators longitudinal studies: limitations and possibilities. *GeoJournal* (2020). https://doi.org/10.1007/s10708-020-10322-0

Malecki, E. (2007). Cities and regions competing in the global economy: knowledge and local development policies. *Environment and Planning C: Government and Policy, 25*, 638-654. https://doi.org/10.1068/c0645

Mallach, A., & Brachman, L. (2013). Regenerating America's legacy cities. *Lincoln Institute of Land Policy.*

Manson, S., Schroeder, J., Van Riper, D., Kugler, T., & Ruggles, S. (2020). *IPUMS National Historical Geographic Information System* (15.0) [Data set]. University of Minnesota. https://www.nhgis.org/

Owen, S.M., MacKenzie, A.R., Stewart, H.E., Donovan, R.G., Hewitt, C.N., Bunce, R.G.H., & Stark, G. (2006). Urban land classification and its uncertainties using principal component and cluster analyses: A case study for the UK West Midlands. *Landscape and Urban Planning, 78*(4), 311-321. https://doi.org/10.1016/j.landurbplan.2005.11.002

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine, Series 6, 2*(11), 559–572.

Schmidtlein, M., Deutsch, R., Piegorsch, M., & Cutter, S. (2008). A Sensitivity Analysis of the Social Vulnerability Index. *Risk Analysis, 28*(4), 1099-1114. https://doi.org/10.1111/j.1539-6924.2008.01072.x

Schofer, E., & Meyer, J. (2005). The Worldwide Expansion of Higher Education in the Twentieth Century. *American Sociological Review, 70*, 898-920. https://doi.org/10.1177%2F000312240507000602

Social Explorer (2021). *1970 Census on 2010 Geographies* [Data set]. Social Explorer. https://www.socialexplorer.com/explore-maps

Starkweather, J. (2010). Principal Components Analysis vs. Factor Analysis…and Appropriate Alternatives. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/Benchmarks/PCAvsFAvsAA_JDS_July2010.pdf

U.S. Census Bureau. (1984, June). *Census Bureau Regions and Divisions with State FIPS Codes.* U.S. Census Bureau. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

U.S. Census Bureau. (1995). *Map of the United States, Showing Census Divisions and Regions*. U.S. Census Bureau. https://www.census.gov/prod/1/gen/95statab/preface.pdf

U.S. Census Bureau. (2020, April 1). *Metropolitan Statistical Areas*. U.S. Census Bureau. https://www.census.gov/programs-surveys/metro-micro/about.html

Van Leuven, A., & Hill, E. (2020). Legacy Regions, Not Legacy Cities: Growth and Decline in City-Centered Regional Economies. *SSRN*. http://dx.doi.org/10.2139/ssrn.3672696

Vey, J. (2007). Restoring prosperity: the state role in revitalizing America's older industrial cities. *The Brookings Institution*.

Vicino, T.J., Hanlon, B., & Short, J.R. (2011). A Typology of Urban Immigrant Neighborhoods. *Urban Geography, 32*(3), 383-405. https://doi.org/10.2747/0272-3638.32.3.383

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems, 2*, 37-52. https://doi.org/10.1016/0169-7439(87)80084-9

Zebardast, E., & Ghanooni, H. (2019). An Analysis of Urban Sprawl Using Factor Analysis Technique (Case: Qazvin City Districts). *Urban Economics and Management, 7*(2(26)), 65-84. https://doi.org/10.1007/s40808-019-00674-z

Zhang, W. (2017). Principal Component Analysis (PCA) in Smart Growth Theory. *Advances in Engineering Research, 114*, 495-498. https://dx.doi.org/10.2991/ammee-17.2017.96

Appendix A - Variable Collection (Literature Review)		
Variable	*Data Source*	*Research studies utilizing variable (consult references for additonal information)*
MSA-wide Population density per square mile	U.S. Census Data	Van Leuven & Hill, 2020
Percent not-foreign born	U.S. Census Data	Van Leuven & Hill, 2020
Percent over age 65	U.S. Census Data	Van Leuven & Hill, 2020; Blumenthal et al., 2008; Schmidtlein et al., 2008
Percent under the age of 18	U.S. Census Data	Van Leuven & Hill, 2020; Blumenthal et al., 2008
Civilian labor-force participation rate	U.S. Census Data	Van Leuven & Hill, 2020; Berube & Murray, 2018; Vey, 2007; Schmidtlein et al., 2008
Number of research-intensive universities per 100,000	U.S. Census Data	Van Leuven & Hill, 2020; Blumenthal et al., 2008
Presence of a state capital	State capital dummy variable	Van Leuven & Hill, 2020; Blumenthal et al., 2008
LQ of MSA's GMP in manufacturing	U.S. Census Data	Van Leuven & Hill, 2020; Blumenthal et al., 2008
Intermodal freight hubs per square mile in MSA	n/a	Van Leuven & Hill, 2020
Number of enplanments per-capita from all airports in MSA	n/a	Van Leuven & Hill, 2020; Blumenthal et al., 2008
Natural log of all buildings designated as historic properties (MSA)	n/a	Van Leuven & Hill, 2020
Percent change in central city's population since peak decennial year	U.S. Census Data	Van Leuven & Hill, 2020
Vacancy of housing units	U.S. Census Data	Van Leuven & Hill, 2020
Median value of owner-occupied housing units	U.S. Census Data	Van Leuven & Hill, 2020; Schmidtlein et al., 2008
Number of property crimes per 1000 people	FBI Uniform Crime Reporting	Van Leuven & Hill, 2020; Mallach & Brachman, 2013
Poverty rate	U.S. Census Data	Van Leuven & Hill, 2020; Vey, 2007; Mallach & Brachman, 2013; Furdell et al., 2005; Schmidtlein et al., 2008
Gini coefficient	U.S. Census Data	Van Leuven & Hill, 2020
Percent of all bridges in MSA deemed "poor" or "structurally deficient"	n/a	Van Leuven & Hill, 2020
Number of "superfund" sites per square mile in MSA	National Priorities List, U.S. Environmental Protection Agency	Van Leuven & Hill, 2020
Percentage of population age 25 and older with bachelor's degree	U.S. Census Data	Van Leuven & Hill, 2020; Mallach & Brachman, 2013; Blumenthal et al., 2008
Percentage of population age 25 and odler with less than high school	U.S. Census Data	Van Leuven & Hill, 2020; Schmidtlein et al., 2008
GDP (gross value added)	U.S. Census Data	Berube & Murray, 2018
Jobs	U.S. Census Data	Berube & Murray, 2018
Jobs at young firms (less than five years old)	U.S. Census Data	Berube & Murray, 2018
GDP per job	U.S. Census Data	Berube & Murray, 2018
Per-capita income	U.S. Census Data	Berube & Murray, 2018; Vey, 2007
Median household income	U.S. Department of Housing and Urban development, U.S. Census Data	Berube & Murray, 2018; Vey, 2007; Furdell et al., 2005; Schmidtlein et al., 2008
Change in employment	U.S. Department of Housing and Urban development, U.S. Census Data	Vey, 2007
Change in annual payrool	U.S. Census Data	Vey, 2007
Change in establishments	U.S. Census Data	Vey, 2007
Unemployment rate	U.S. Department of Housing and Urban development, U.S. Census Data	Vey, 2007; Mallach & Brachman, 2013; Furdell et al., 2005; Schmidtlein et al., 2008

Percent foreign born population	U.S. Census Data	Mallach & Brachman, 2013
Population loss from peak to 2010	U.S. Census Data	Mallach & Brachman, 2013
Population change	U.S. Census Data	Mallach & Brachman, 2013
Household dependency ratio	U.S. Census Data; Brookings calculation	Mallach & Brachman, 2013
Change in median housing price	PolicyMap, Boxwood Means	Mallach & Brachman, 2013
Mortgage ratio	PolicyMap, Boxwood Means, Home Mortgage Disclosure Act; Brookings calculations	Mallach & Brachman, 2013
Graduate students as percentage of city population	Greater Ohio Policy Center, Field Survey (2012)	Mallach & Brachman, 2013
Total research funding	Lombardi, Phillips, Abbey, and Craig (2011)	Mallach & Brachman, 2013
Change in number of jobs	U.S. Census Data	Mallach & Brachman, 2013
Change in population over preceding decade	U.S. Census Data	Furdell et al., 2005
Location Quotient for FIRE	U.S. Census Data	Blumenthal et al., 2008
Black non-Hispanic residents	U.S. Census Data	Blumenthal et al., 2008
Average wage	Bureau of Economic Analysis	Blumenthal et al., 2008
July temperature	countrystudies.us	Blumenthal et al., 2008
Right-to-work state	National Right-to-Work Legal Defense Foundation	Blumenthal et al., 2008
Regions	Regional dummy variables	Blumenthal et al., 2008
Percentage of female participation in civilian labor force	U.S. Census Data	Schmidtlein et al., 2008
Median gross rent for renter-occupied housing units	U.S. Census Data	Schmidtlein et al., 2008
Percentage of population under 5 years	U.S. Census Data	Schmidtlein et al., 2008
Percentage of institutionalized elderly population	U.S. Census Data	Schmidtlein et al., 2008
Average number of people per household	U.S. Census Data	Schmidtlein et al., 2008
Percentage of renter-occupied housing units	U.S. Census Data	Schmidtlein et al., 2008
Percentage of rural farm occupation	U.S. Census Data	Schmidtlein et al., 2008
Percentage of employment in primary industry: farming, fishing, mining or forestry	U.S. Census Data	Schmidtlein et al., 2008
Percentage of Asian or Pacific Islander	U.S. Census Data	Schmidtlein et al., 2008
Percentage of Hispanic persons	U.S. Census Data	Schmidtlein et al., 2008
Percentage employed in transportation, communication, or other public utilities	U.S. Census Data	Schmidtlein et al., 2008
Percentage of population living in urban areas	U.S. Census Data	Schmidtlein et al., 2008
Percentage employed in service occupations	U.S. Census Data	Schmidtlein et al., 2008
Percentage of females	U.S. Census Data	Schmidtlein et al., 2008
Percentage of households that receive Social Security benefits	U.S. Census Data	Schmidtlein et al., 2008

Appendix B - Final Variable List		
Variable	*Data Source*	*Research studies utilizing variable*
MSA-wide Population density per square mile	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020
Percent over age 65	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020; Blumenthal et al., 2008; Schmidtlein et al., 2008
Percent under the age of 18	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020; Blumenthal et al., 2008
Civilian labor-force participation rate	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020; Berube & Murray, 2018; Vey, 2007; Schmidtlein et al., 2008
Vacancy of housing units	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020
Median value of owner-occupied housing units	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Van Leuven & Hill, 2020; Schmidtlein et al., 2008
Poverty rate	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Van Leuven & Hill, 2020; Vey, 2007; Mallach & Brachman, 2013; Furdell et al., 2005; Schmidtlein et al., 2008
Percentage of population age 25 and older with bachelor's degree	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Van Leuven & Hill, 2020; Mallach & Brachman, 2013; Blumenthal et al., 2008
Percentage of population age 25 and odler with less than high school	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Van Leuven & Hill, 2020; Schmidtlein et al., 2008
Jobs	Bureau of Economic Analysis: Regional Economic Information System	Berube & Murray, 2018
Per-capita income	Bureau of Economic Analysis: Regional Economic Information System	Berube & Murray, 2018; Vey, 2007
Median household income	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Berube & Murray, 2018; Vey, 2007; Furdell et al., 2005; Schmidtlein et al., 2008
Unemployment rate	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Vey, 2007; Mallach & Brachman, 2013; Furdell et al., 2005; Schmidtlein et al., 2008
Percent foreign born population	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Mallach & Brachman, 2013
Percentage of Black non-Hispanic residents	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Blumenthal et al., 2008
Percentage of female participation in civilian labor force	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008
Median gross rent for renter-occupied housing units	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of population under 5 years	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of institutionalized elderly population	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Average number of people per household	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of renter-occupied housing units	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of employment in primary industry: farming, fishing, mining or forestry	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of Hispanic persons	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Percentage employed in transportation, communication, or other public utilities	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of population living in urban areas	National Historic Geographic Information System; 1970 U.S. Census; 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of females	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008
Percentage of female-headed households	www.socialexplorer.com; 1970 U.S. Census, 2010 U.S. Census	Schmidtlein et al., 2008

Appendix C

Methodology Considerations

Four tests were performed on the data set from 1970 and 2010 before they were analyzed: checking for a linear relationship between variables, checking for sampling adequacy (KMO Measure of Sampling Adequacy), checking that the data is suitable for reduction (Bartlett's test of sphericity), and checking that there are no significant outliers. SPSS was used to perform each test on the data. The software's Linear Regression tool was used to check for a linear relationship between variables. Multiple sets of variables were analyzed during this test. The Factor Analysis tool was utilized to perform both the KMO Measure of Sampling Adequacy and Bartlett's test of sphericity. The Explore tool printed outliers for each variable. As described in the Analysis section, for data set from 1970, there was a linear relationship between variables, the KMO measure was 0.673, Bartlett's test yielded a value of < 0.001, and significant outliers were then removed from the data set. For the data set from 2010, there was a linear relationship between variables, and significant outliers were removed, but the KMO measure and Bartlett's test measure did not have an output. The PCA still outputted from this dataset, but this means that there are linear dependencies between variables.

Performing Principal Component Analysis with an unrotated component matrix was attempted, but no component for either PCA on the Component Correlation Matrix had a value above 0.32. A Varimax rotation was chosen next, and several components for both PCAs had a value above 0.32. Components accounting for approximately 80% of each data set were chosen, resulting in eight components for the first PCA and six components for the second. For the first PCA's Rotated Component Matrix, values below 0.45 for each variable were suppressed. For the second PCA's Rotated Component Matrix, values below 0.5 were suppressed. These

suppressions enable most variables to be strongly correlated with one component for each PCA. Some variables (like MSA-Wide Population Density per square mile) were present for several components.

To calculate component scores for each MSA for both data sets, the variable loadings were multiplied by the standardized z-scores for every variable within each principal component and summing the totals. This process was completed within Microsoft Excel for both data sets.